The computation of the regression coefficient for the dataset shown in Table 2.2 will be the topic of an exercise. But let us assume that we have computed the value of $\hat{b}$, then the predicted value of the body weight $\widehat{y_s}$ for an animal s is computed based on the measured breast circumference x_s of animal s as follows

$$\widehat{y_s} = \hat{b} * x_s \tag{2.22}$$

It has to be noted that the prediction $\widehat{y_s}$ is only valid, if the measured value x_s is close to the measured predictors that were used to estimate $\hat{b}$. For our example with body weight and breast circumference, we could not use the same regression line to predict the body weight for calves, if $\hat{b}$ was estimated with data of adult bulls.

2.9 Regression On Dummy Variables

In a regression model (such as shown in (2.10)) both the response variable and the predictor variables are continuous variables. Examples of such variables are `body weight` and `breast circumference` which are both measured and the measurements are expressed as real numbers. In contrast to such a regression model, the statistical model shown in (2.6) has a continuous response, but the predictor variables are discrete variables. The predictor variables are assumed to be genotypes of a certain set of SNP genotypes and hence these genotypes can only have a fixed number of states. Under the assumption of bi-allelic Loci, a SNP locus can have just three genotypes and hence the predictor variable that is used to represent any given SNP-locus can only take three discrete states.

Figure 2.3 shows the difference between a regression model as the one of `body weight` on `breast circumference` and a fixed linear effects model where one locus has an effect on a quantitative trait. In the left diagram of Figure 2.3 the red line denotes the regression line. This line is meaningful because on the x-axis and on the y-axis every single point of the red line would be valid observations. On the x-axis of the diagram on the righthand side, only three values are possible. In the diagram they are shown as Genotypes G_1G_1, G_1G_2 and G_2G_2. We will see very soon that in our statistical model, they will be encoded by 1, 0 and -1. The response variable in the diagram on the right of Figure 2.3 is a continuous random variable, similarly to the regression model shown in the left diagram. This combination of continuous response variable on a discrete type of variable lead to the term **regression on dummy variables** because the predictor variables are not continuous but just discrete levels of a certain factor. In this lecture, we are using **fixed linear effects model** rather than regression on dummy variables for the same type of model. The term of fixed linear effects model was used, because in the next chapter in Genomic BLUP we are going to introduce mixed linear effects model which are an extension of the fixed linear effects model used in this chapter.

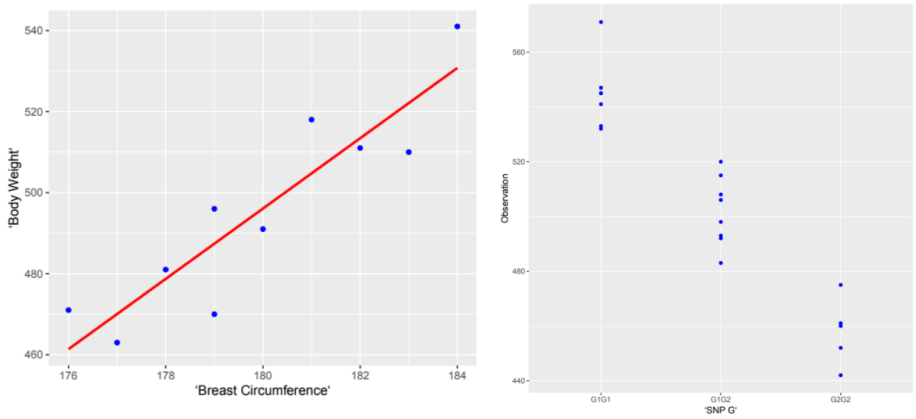

Figure 2.3: Comparison Between Regression Model And Fixed Linear Effects Model With An SNP-Locus As A Discrete Predictor Variables

2.9.1 Fixed Linear Effects Model For SNP Data

We are using genetic data and assume that the SNP genotypes have an effect on a quantitative trait. Our goal is to predict genomic breeding values based on the information from the SNP genotypes for the quantitative traits. We have seen that under some simplifying assumptions of additivity of the genetic effects, the genomic breeding values depend on the absolute value of the genotypic values (a values) of the homozygous SNP genotypes. Hence all we need to know from our analysis of the data under a fixed linear effects model are the a values for each SNP locus. The decomposition of the phenotypic observation shown in 2.4.1 under the assumed genetic model tells us that the phenotypic observation can be explained as a linear function of the genotypic values of the SNP genotypes plus a random error term. The fact that our genetic model is a fixed linear effects model that uses phenotypic observations as response and SNP loci as predictors allows us to set up the following model for an example data set shown in the following subsection.

2.9.2 Example Data Set With SNP Loci And A Phenotypic Observation

We are using the dataset shown in Table 2.3 as an example on how to use a fixed linear effects model to estimate the genotypic value of the SNP genotypes.

Instead of fitting individual effects for the different SNP genotypes to explain the response variable, we are directly including the genotypic values a_G and a_H into the fixed effects linear model. How the genotypic values are related to the SNP genotypes is also shown in Table 2.3. For all animals in Table 2.3, we can

Table 2.3: Animals With A Single SNP Locus Affecting A Quantitative Trait

Animal	SNP G	Genotypic Value G	SNP H	Genotypic Value H	Observation
1	G_1G_1	a_G	H_1H_2	0	510
2	G_1G_2	0	H_1H_1	a_H	528
3	G_1G_2	0	H_1H_1	a_H	505
4	G_1G_1	a_G	H_2H_2	$-a_H$	539
5	G_1G_1	a_G	H_1H_1	a_H	530
6	G_1G_2	0	H_1H_2	0	489
7	G_1G_2	0	H_2H_2	$-a_H$	486
8	G_2G_2	$-a_G$	H_1H_1	a_H	485
9	G_1G_2	0	H_2H_2	$-a_H$	478
10	G_2G_2	$-a_G$	H_1H_2	0	479
11	G_1G_1	a_G	H_1H_2	0	520
12	G_1G_1	a_G	H_1H_1	a_H	521
13	G_2G_2	$-a_G$	H_1H_2	0	473
14	G_2G_2	$-a_G$	H_1H_2	0	457
15	G_1G_2	0	H_1H_1	a_H	497
16	G_1G_2	0	H_1H_2	0	516
17	G_1G_1	a_G	H_1H_2	0	524
18	G_1G_1	a_G	H_1H_2	0	502
19	G_1G_1	a_G	H_2H_2	$-a_H$	508
20	G_1G_2	0	H_1H_2	0	506

write the model equations in matrix-vector notation as

$$y = Xb + \epsilon \tag{2.23}$$

where y is the vector of observations, b is a vector of genotypic values plus an intercept, X is a design matrix linking the elements in b to y and ϵ is a vector of random errors. Writing out the matrices and vectors leads to

$$\begin{bmatrix} 510 \\ 528 \\ 505 \\ 539 \\ 530 \\ 489 \\ 486 \\ 485 \\ 478 \\ 479 \\ 520 \\ 521 \\ 473 \\ 457 \\ 497 \\ 516 \\ 524 \\ 502 \\ 508 \\ 506 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_0 \\ a_G \\ a_H \end{bmatrix} + \epsilon \tag{2.24}$$

2.9.3 Parameter Estimation In A Fixed Linear Effects Model

The goal for model (2.23) is to get an estimate for the unknown parameters b_0, a_G and a_H. In section 2.9.3 we saw how unknown parameters can be estimated for a regression model using least squares. When applying the least squares method, we did not make any assumptions about the predictor variables. The minimization of the sum of the squared residuals can also be applied for the fixed linear effects model. This minimization leads to the same normal equations

$$X^T X b^{(0)} = X^T y \tag{2.25}$$

So far everything was identical to the case of the regression model. But when trying to find a solution for (2.25) we have to account for the different nature of the design matrix X. In the regression model this matrix X contains real numbers. In our example of a fixed linear effects model, the matrix X just contains just the three number -1, 0 and 1[4]. The fact that the matrix X contains only a few discrete values makes it very likely that X does not have full column rank. That means it is very likely that some columns of X can be expressed as linear combinations of other columns. This linear dependence of the columns of X causes the matrix $X^T X$ to be singular and hence the inverse of

[4]In most other fixed linear effects models, the design matrix contains just 0 and 1.

X^TX cannot be computed. Whenever the matrix X^TX is singular, the solution given in (2.17) cannot be computed.

The normal equations in (2.25) are written with the symbol $b^{(0)}$ to denote that the equations do not have a single solution $b^{(0)}$ in the sense that we were able to compute them in the case of the regression model. In the case where X^TX is singular, there are infinitely many solutions $b^{(0)}$. These solutions can be expressed as

$$b^{(0)} = (X^TX)^-X^Ty \tag{2.26}$$

where $(X^TX)^-$ denotes the **generalized inverse** of the matrix X^TX. A generalized inverse G of a given matrix A is defined as the matrix that satisfies the equation $AGA = A$. The matrix G is not unique. Applying the concept of a generalized inverse to a system of equations $Ax = y$, it can be shown that $x = Gy$ is a solution, if G is a generalized inverse of A. Because G is not unique, there are infinitely many solutions corresponding to $\tilde{x} = Gy + (GA - I)z$ where z can be an arbitrary vector of consistent order. Applying these statements concerning generalized inverses and solutions to systems of equations to (2.26), it means that $b^{(0)}$ is not a unique solution to (2.25) because the generalized inverse $(X^TX)^-$ is not unique. As a consequence of that the solution $b^{(0)}$ cannot be used as an estimate of the unknown parameter vector b.

The numeric solution of the analysis of the example dataset given in Table 2.3 is the topic of an exercise. When developing that solution, we will see that some linear functions of $b^{(0)}$ can be found which do not depend on the choice of the generalized inverse $(X^TX)^-$. Such functions are called **estimable functions** and can be used as estimates for the unknown parameter vector b. More details about generalized inverses and estimable functions can be found in [Searle, 1971].